The transcriptional foundation of pluripotency

Development 136, 2311-2322 (2009) doi:10.1242/dev.024398

The transcriptional foundation of pluripotency

Ian Chambers and Simon R. Tomlinson

A fundamental goal in biology is to understand the molecular basis of cell identity. Pluripotent embryonic stem (ES) cell identity is governed by a set of transcription factors centred on the triumvirate of Oct4, Sox2 and Nanog. These proteins often bind to closely localised genomic sites. Recent studies have identified additional transcriptional modulators that bind to chromatin near sites occupied by Oct4, Sox2 and Nanog. This suggests that the combinatorial control of gene transcription might be fundamental to the ES cell state. Here we discuss how these observations advance our understanding of the transcription factor network that controls pluripotent identity and highlight unresolved issues that arise from these studies.

Introduction

Pluripotency is the capacity of a cell to give rise to differentiated derivatives that represent each of the three primary germ layers. Pluripotency is a property of the cells that are located within the inner cell mass (ICM) of the developing blastocyst. These cells can be explanted and embryonic stem (ES) cell lines established from them that can be cultured in vitro, essentially indefinitely. The maintenance of pluripotent identity requires either the stimulation of mouse ES cells by leukemia inhibitory factor (LIF) and by bone morphogenetic protein (BMP) or the treatment of cells with a cocktail of enzyme inhibitors, which are thought to block the action of pro-differentiative signals generated autonomously by ES cells (for reviews, see Chambers and Smith, 2004; Silva and Smith, 2008). ES cells possess many features that make them attractive for studying the molecular basis of cell identity (Box 1). Functional studies have been of central importance in identifying a group of transcription factors that affect the pluripotent identity of ES cells (Chambers et al., 2003; Chambers et al., 2007; Ema et al., 2008; Fujikura et al., 2002; Masui et al., 2007; Niwa et al., 2000; Niwa et al., 2005). Within this group, the transcription factors Oct4 (Pou5f1), Sox2 and Nanog are crucial for the efficient maintenance of ES cell identity (Chambers et al., 2007; Masui et al., 2007; Niwa et al., 2000). During mouse development, the specification of pluripotent cell identity requires the embryonic genome to express *Oct4* (Nichols et al., 1998) and *Nanog* (Mitsui et al., 2003), but not *Sox2* (Avilion et al., 2003), perhaps owing to the presence of long-lived maternal Sox2 protein (Avilion et al., 2003).

Genome-wide studies have highlighted the colocalisation of Oct4, Sox2 and Nanog in ES cell chromatin and the existence of a transcriptional network that might direct ES cell identity (Boyer et al., 2005; Loh et al., 2006). Subsequent studies have identified multiple additional transcription factors that colocalise with Oct4, Sox2 and Nanog (Chen et al., 2008b; Cole et al., 2008; Kim et al., 2008) and have connected miRNA-encoding genes to this circuitry (Marson et al., 2008). Such studies have expanded our view of the transcription factor network that controls the ES cell state and have generated a wealth of bioinformatic data for further analysis.

Here we review the experiments that have informed our understanding of the biological function of Oct4, Sox2 and Nanog in ES cell self-renewal. We discuss experiments that have identified proteins that interact with these molecules, before addressing the findings of global chromatin localisation studies. The latter have in part been informed by parallel studies on the induction of pluripotency in somatic cells [for recent reviews on induced pluripotency, we refer the reader to the literature (Hochedlinger and Plath, 2009; Jaenisch and Young, 2008; Yamanaka, 2007)]. Finally, we examine unanswered questions that these studies have raised.

DNA-binding characteristics of Nanog, Oct4 and Sox2

The precise manner in which Nanog, Oct4 and Sox2 interact with DNA is likely to make a major contribution to their function. Nanog has a single homeodomain that binds to DNA (Fig. 1A) (Jauch et al., 2008). The Nanog homeodomain is a variant; although most closely related to the Nkx family, Nanog is not part of this family. The level of identity between the Nanog homeodomain and other homeodomains is below that required for assignment to a particular family (Kappen et al., 1993). Moreover, sequences that are conserved among the Nkx family members are absent from Nanog (Lints et al., 1993). The DNA sequence bound by Nanog is a matter of some controversy. The in vitro selection of random oligonucleotides by recombinant Nanog expressed in *E. coli* has suggested that Nanog binds to the homeodomain core recognition sequence TAAT (Mitsui et al., 2003). More detailed DNA-binding analysis of the purified homeodomain alone has suggested that Nanog binds to the extended sequence TAATGG (Jauch et al., 2008). By contrast, results from global localisation studies in ES cells have suggested that Nanog binds to the sequence CATT (Loh et al., 2006).

Oct4 belongs to the Octamer class of transcription factors that recognise an 8-bp DNA site (hence the name) with the consensus ATGCAAAT (Falkner and Zachau, 1984; Parslow et al., 1984). Together with Pit and Unc proteins, Oct proteins define the POU (Pit, Oct and Unc) class of transcription factors that interact with DNA through two DNA-binding domains (Fig. 1B): a low-affinity

Box 1. The utility of ES cells in studying the molecular basis of cell identity

ES cells are defined by the simultaneous possession of the seemingly incongruent properties of self-renewal and the capacity to differentiate into cellular derivatives of all primary germ layers.

ES cells can be cultivated in large numbers, facilitating their biochemical analysis.

ES cells can be readily modified genetically, allowing the effects on cell identity of altering the intracellular environment to be monitored.

ES cell identity can also be altered by changing the culture conditions, the effects of which can be monitored and assessed.

MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, University of Edinburgh, King's Buildings, Edinburgh EH9 3JQ, UK.

E-mails: ichambers@ed.ac.uk; simon.tomlinson@ed.ac.uk

POU-specific domain (POU_S) and a higher affinity homeodomain (POU_{HD}) (Klemm and Pabo, 1996). Each POU domain contacts 4 bp in the major groove of the cognate DNA site, thereby placing each DNA-binding domain on either side of the helix and effectively encircling the DNA (Phillips and Luisi, 2000).

Sox2 is a member of a superfamily of proteins that all possess a High mobility group (HMG) box DNA-binding domain (Fig. 1C). Sox subfamily members are defined by the relationship of their HMG box to that of the testes determining factor Sry, the archetypal Sox protein from which the family takes its name (Sox, Sry HMG box) (Bowles et al., 2000). The Sox2 HMG domain interacts with DNA through the minor groove of the consensus sequence $^{A}/_{T}{}^{A}/_{T}$CAAAG (Bowles et al., 2000). Interestingly, a methionine residue within the Sox2 HMG domain intercalates between bases in the binding site and acts like a cantilever to cause DNA bending (Weiss, 2001).

Oct4 and Sox2 bind DNA co-operatively (Ambrosetti et al., 1997; Ambrosetti et al., 2000). It is noteworthy that at validated target sites, the non-palindromic Oct4 and Sox2 cognate sequences always occur adjacent to one another in a particular relative orientation. Furthermore, the interaction of the POU_S domain with the major groove and of the HMG domain with the minor groove positions these two DNA-binding domains adjacent to one another on the same side of the helix (Fig. 2). Although structural studies of Oct4 and Sox2 bound to DNA have not been performed, structural analyses of the DNA-binding domains of Oct1 (Pou2f1) and Sox2 complexed with DNA (Remenyi et al., 2003; Williams et al., 2004) suggest how Sox2 and Oct4 might interact with DNA (Fig. 2). Interestingly, analysis of the Oct-DNA binary complex has revealed the existence of two Oct/DNA conformations, one in which both DNA-binding domains contact the DNA, and one in which only the higher affinity POU_{HD} contacts the DNA (Williams et al., 2004). By contrast, in the Oct-Sox-DNA ternary complex, the POU_S domain was always in contact with the DNA. This is due to side chains on the Sox2 HMG domain and the Oct POU_S domain, which interact with one another and hold the POU_S domain onto the DNA. This study not only offers an explanation for the relatively inflexible spatial relationship observed between the non-palindromic DNA recognition sequences for Oct4 and Sox2 at validated targets, but also provides an indication of how protein interactions might begin to build up a stable, multi-protein machine for the combinatorial control of gene transcription.

In marked contrast to these structural studies, which aim to provide a view of the transcriptional machinery at atomic resolution, a recent trend has been to examine the localisation of proteins throughout chromatin as a means of elucidating the transcriptional control of cell identity. We discuss these studies in more detail below. First, we discuss the phenotypic effects on pluripotency of altering the levels of Nanog, Oct4 and Sox2.

Modulation of Oct4, Sox2 and Nanog

In 2000, a bench-mark study defined the effects on ES cell self-renewal of altering the dose of a transcription factor, Oct4 (Niwa et al., 2000). In this study, the self-renewal of *Oct4*-null mouse ES cells was sustained by the expression of a tetracycline-suppressible *Oct4* transgene. Upon silencing of *Oct4* with tetracycline, ES cells differentiated into trophectoderm, the lineage that supplies trophoblast cells for the developing placenta (Kunath et al., 2004). Using the same tetracycline-suppressible transgene in *Oct4* heterozygous ES cells, the additional, surprising observation was made that increasing Oct4 protein levels above ~150% of that in wild-type ES cells caused differentiation into a mixed population of cells that express both mesoderm and endoderm markers (Niwa et al., 2000).

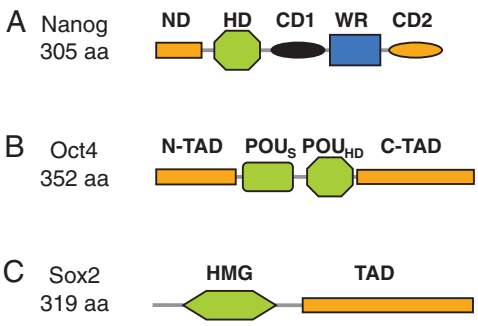

Fig. 1. Nanog, Oct4 and Sox2 protein domains. Each protein is divided into domains, either real or putative. DNA-binding domains are shown in green and regions with reported trans-activating potential in orange. (**A**) Nanog can be divided into N-terminal and C-terminal halves. The N-terminal half contains a DNA-binding homeodomain (HD) and an N-terminal domain (ND). The C-terminal half contains a dimerisation domain (blue) referred to as the tryptophan repeat (WR), in which every fifth residue is a tryptophan (Mullin et al., 2008), that separates C-terminal domain 1 (CD1) from C-terminal domain 2 (CD2). (**B**) Oct4 has DNA-binding domains comprising a POU-specific DNA-binding domain (POU_S) and a POU-homeodomain (POU_{HD}), each of which can interact independently with DNA (Herr and Cleary, 1995), as well as transactivation domains located N-terminal (N-TAD) or C-terminal (C-TAD) to the POU domain. (**C**) Sox2 is a High mobility group (HMG) family member and has a single HMG DNA-binding domain and a transactivation domain (TAD). The size of each protein is indicated in amino acid residues (aa). Drawings are not to scale.

Deletion of *Sox2* in ES cells causes a trophectodermal differentiation phenotype similar to that seen following the deletion of *Oct4* (Masui et al., 2007). A major surprise from these studies (Masui et al., 2007) was the discovery that the expression of many Oct/Sox target genes was not greatly affected by the loss of *Sox2*. This is probably a consequence of the expression of the additional Sox family members Sox4, Sox11 and Sox15 in ES cells. The key distinguishing contribution of Sox2 appears to be to maintain *Oct4* expression: consistent with this view is the finding that the enforced expression of *Oct4* can rescue ES cells from differentiation induced by the loss of *Sox2* (Masui et al., 2007). This has been proposed to reflect the regulation by Sox2 of the genes that encode the transcription factors Nr5a2 (nuclear receptor subfamily 5, group A, member 2; also known as Lrh1) and Nr2f2 (nuclear receptor subfamily 2, group F, member 2; also known as CoupTFII), which in turn act on *Oct4*. In addition to these knockout studies, Sox2 overexpression studies have shown that under conditions that favour differentiation, Sox2-overexpressing ES cells are biased towards neural differentiation (Kopp et al., 2008; Zhao et al., 2004). However, whereas a recent study found that under self-renewing conditions Sox2-overexpressing cells differentiate (Kopp et al., 2008), this was not observed in an earlier study (Zhao et al., 2004).

Nanog was isolated simultaneously by two groups. In one study, in silico analysis identified ES cell-specific transcripts, including *Nanog* (Mitsui et al., 2003). In our laboratory, expression cloning from an ES cell cDNA library identified Nanog as a molecule capable of conferring cytokine-independent self-renewal upon transfected ES cells (Chambers et al., 2003). Subsequent analyses have demonstrated that reduction in the level of *Nanog* increases the

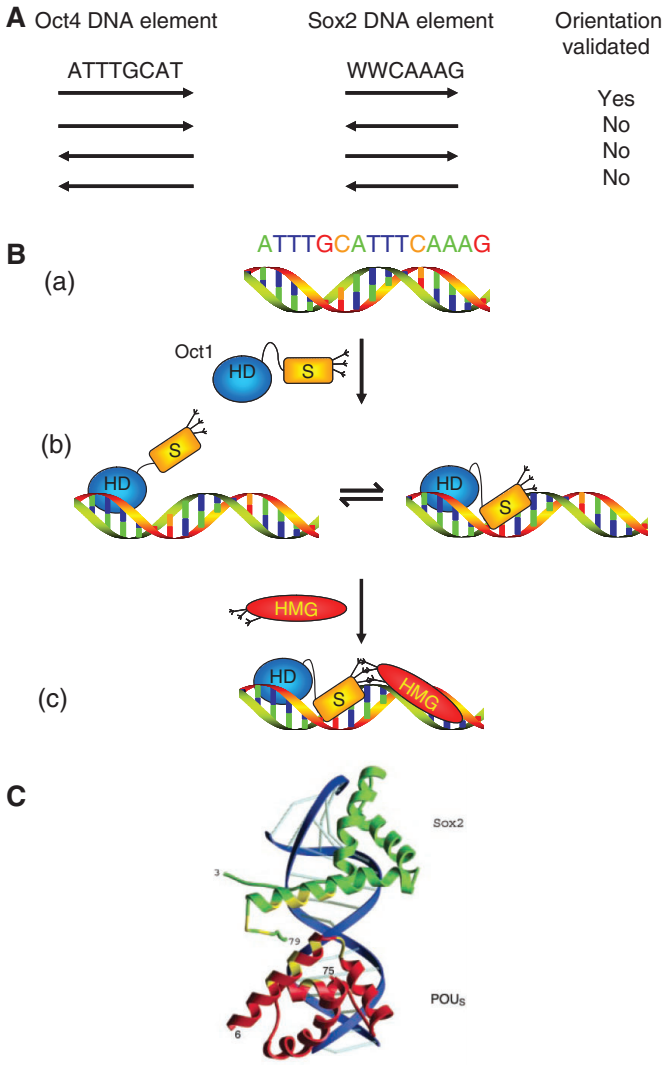

Fig. 2. The Oct/Sox DNA motif and the ternary structure of Oct-Sox-DNA. (**A**) The DNA sites recognised by Oct4 and Sox2 are both non-palindromic and might be expected to exist in four distinct relative orientations. However, one of these orientations predominates at sites that have been validated by reporter assays. (**B**) Nuclear magnetic resonance (NMR) analysis of Oct1 bound to DNA shows two conformations. (**a**) The sequence of a composite Oct/Sox site is illustrated. (**b**) After binding of Oct1 to the DNA, the binary complex exists in two conformations; the POU homeodomain (HD) contacts DNA directly in both, but in only one is the POU-specific DNA-binding domain (S) also in direct contact with DNA. (**c**) DNA binding by Sox2 via its HMG DNA-binding domain (HMG) provides stabilising side-chain interactions that lock POU_S onto the DNA. (**C**) The ternary structure of Oct1 (red) and Sox2 (green) bound to the *Hoxb1* regulatory element (blue DNA). The backbone position of residues, the side chains of which provide stabilising interactions, are highlighted (yellow). Reprinted with permission from Williams et al. (Williams et al., 2004).

propensity of cells to differentiate (Chambers et al., 2007; Hatano et al., 2005; Ivanova et al., 2006). However, *Nanog*-null ES cells continue to self-renew, indicating that the loss of Nanog does not commit ES cells to differentiation (Chambers et al., 2007). Although these studies tell us about the consequences for pluripotency of altering the level of key transcription factors, they do not tell us how such molecules work.

Identifying Oct4 and Nanog interaction partners

One approach to the question of how transcriptional regulators function is to use affinity-based methods in combination with mass spectrometry to identify the proteins with which such regulators interact. As we discuss below, this approach has identified proteins that interact with the core pluripotency factors.

In one study, antibodies against genetically unmodified Nanog protein identified an interaction between Nanog and Smad1 (a signal transducer in the BMP signalling pathway) (Suzuki et al., 2006). In addition, an epitope-tagging approach (Fig. 3A), in which a Flag affinity peptide was fused to the N-terminus of Nanog, allowed immobilised anti-Flag immunoglobulins to be used to purify Nanog-containing complexes. Subsequent fractionation of interacting proteins, followed by trypsinisation and mass spectrometric analysis of peptides (Fig. 3) identified Sall4 (sal-like 4, a zinc-finger transcription factor of the Spalt family) as a Nanog-interacting protein (Wu et al., 2006). A conceptually similar approach based on the biotinylation of transgenic constructs that encode a protein of interest was used to identify multiple proteins that interact with Nanog and Oct4 (Wang et al., 2006). This method (Fig. 3B), developed by Strouboulis and colleagues (de Boer et al., 2003), uses ES cells that express a biotin ligase encoded by the *E. coli BirA* gene (Driegen et al., 2005). Stable transfection of these cells with expression constructs that contain an additional 16 residues appended to the open reading frame makes the encoded protein a ligase substrate and a biotin adduct is added to a lysine residue within the tag. In order to minimally perturb the transcription factor network, ES cell clones expressing the additive transgenes at levels below those of the endogenous protein were examined (Wang et al., 2006). In this way, the interaction of Nanog with multiple transcription regulators was detected, including: Sall4; Nr0b1 (nuclear receptor subfamily 0, group B, member 1); Nac1 (nucleus accumbens associated 1; also known as Nacc1); Esrrb (estrogen-related receptor beta; also known as Nr3b2); Zfp281 (zinc-finger protein 281); Hdac2 (histone deacetylase 2) and Sp1. Oct4 was shown to interact with an overlapping set of proteins, including Sall4, Hdac2 and Sp1. Moreover an interaction between Nanog and Oct4 detected by mass spectrometry was confirmed in this study by co-immunoprecipitation from transfected cell lysates by epitope tagging (Wang et al., 2006). A separate study used a bacterially expressed glutathione S-transferase-Oct4 fusion protein to pull down in vitro translated Nanog with glutathione-conjugated beads (Liang et al., 2008). These studies are illustrative of an on-going debate regarding the relative merits of the different methods for determining protein-protein interactions (Chatr-Aryamontri et al., 2008; Mackay et al., 2007; Welch, 2009). Briefly, the concern is that protein interaction data generated by affinity-based proteomic approaches of the type outlined above are often validated by co-immunoprecipitation from cell lysates. As conditions for co-immunoprecipitation can be varied to 'optimise' the detection of an interaction (Mackay et al., 2007), false-positive protein-protein interactions might be reported. Computational methods can be used to set thresholds to eliminate false positives (Chatr-Aryamontri et al., 2008). In the case of the Nanog-Oct4 interaction, despite the fact that an interaction between endogenous Oct4 and Nanog has yet to be confirmed, multiple approaches have provided support for their interaction that exceeds that available for any of the other protein-protein interactions of the pluripotency network factors, other than Oct4 and Sox2. Ultimately, biophysical methods will be required to rigorously characterise this and other interactions.

The opposite problem, i.e. false negativity, appears to be represented by the case of Smad1, which was not detected by mass spectrometry (Liang et al., 2008; Wang et al., 2006) but did co-

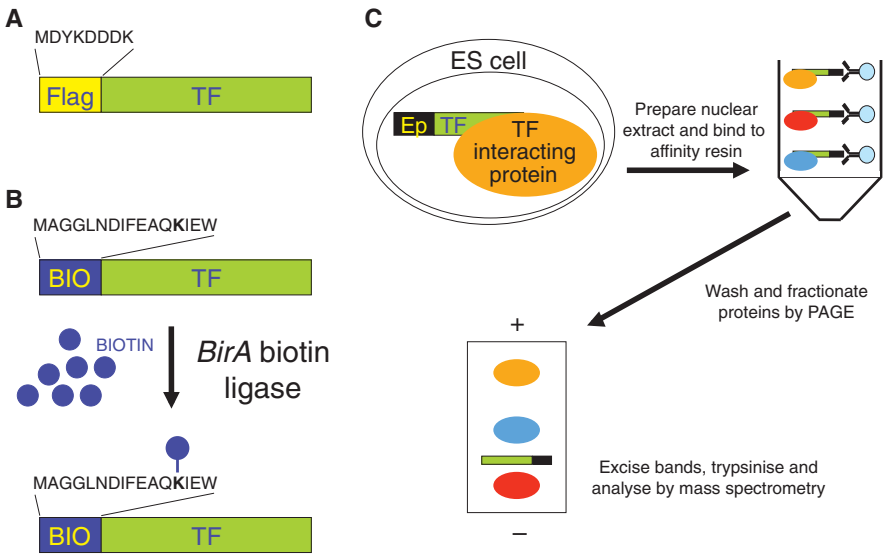

Fig. 3. Affinity-based methods for identifying interacting proteins. (**A**,**B**) Modification of a transcription factor (TF) using epitope tags. (A) The TF open reading frame is modified by in-frame fusion to an epitope tag encoding the so-called Flag peptide against which antibodies are commercially available. (B) A 16-residue tag (BIO) added to the TF is a substrate for biotinylation when expressed in cells that are engineered to express the *E. coli BirA* gene, which encodes a biotin ligase. (**C**) Affinity purification and partner identification strategy. Nuclear extracts are prepared from ES cells that express an epitope-tagged (Ep) form of a TF under conditions in which binding of TF-interacting proteins to the TF is maintained. Protein complexes are then incubated with an affinity reagent (anti-Flag IgG or streptavidin) that is immobilised on a solid support. After washing to remove contaminating proteins, the captured proteins are separated by electrophoresis, excised from the gel, digested with protease and peptides identified by mass spectrometry. PAGE, polyacrylamide gel electrophoresis.

immunoprecipitate with Nanog (Suzuki et al., 2006). Such cases can arise when the protein is poorly digested prior to mass spectrometry, for instance because of a paucity of trypsin cleavage sites.

An additional screen for Nanog-interacting proteins was conducted using an anti-Nanog antibody for affinity purification from ES cells. Many of the proteins previously identified by Wang et al. (Wang et al., 2006) were also detected in this study, including Hdac2 (Liang et al., 2008). Interestingly, although Hdac2 is a component of the nucleosome remodelling and histone deacetylation (NuRD) complex (McDonel et al., 2009), several proteins usually found in NuRD, including Rbbp7 (retinoblastoma binding protein 7) and Mbd3 (methyl-CpG binding domain protein 3), were suggested to be present at substoichiometric levels relative to the remaining NuRD components (Liang et al., 2008). This might indicate that a particular and unusual form of NuRD associates with Oct4 and Nanog, as the authors claim, or could reflect distinct detection sensitivities for different NuRD subunits.

The affinity-based approaches summarised above have been important in defining additional components of the transcriptional machinery that Oct4 and Nanog interact with, and in several cases iterative epitope tagging and affinity purification have further extended the protein interaction network (Wang et al., 2006). Much remains to be discovered about the mechanisms by which partners interact. An important question concerns the significance of these interactions in terms of pluripotent gene regulation, an issue we turn to below.

Global localisation of pluripotency regulators to chromatin

The list of Oct4 interactors identified in the biotin-based affinity study (Wang et al., 2006) lacked one of the best-studied partners of Oct4, namely Sox2. This might be because Oct4 and Sox2 interact most stably when bound to adjacent sites on DNA. It is therefore pertinent to ask which transcriptional regulators colocalise when bound to DNA. In reality, this is hard to achieve at high resolution on a global scale. However, initial studies have shown that OCT4, SOX2 and NANOG frequently bind to sites that are closely localised in the chromatin of human ES cells (Boyer et al., 2005). Furthermore, Oct4 and Nanog have been found at closely localised sites in mouse ES cells (Loh et al., 2006). Global colocalisation of Oct4 and Sox2 might have been expected given the known co-operative DNA binding of Oct4 and Sox2 (Ambrosetti et al., 1997). However, there were no experimental data to predict the close localisation of Nanog with Oct4 and Sox2. Subsequent large-scale chromatin immunoprecipitation studies have shown that the close localisation of Nanog with Oct4 and Sox2 is not peculiar to Nanog but is shared with several other pluripotency transcription factors, including Klf4 (Kruppel-like factor 4), Esrrb and Tcf4 (T cell factor 4) (Chen et al., 2008b; Cole et al., 2008; Kim et al., 2008).

Large-scale global localisation studies generally follow protocols that consist of two stages. In the first stage, chromatin associated with a protein of interest is purified using an affinity-based approach akin to the techniques discussed in the previous section; this step is generally referred to as chromatin immunoprecipitation (ChIP). In the next stage, the DNA associated with the protein of interest is purified and analysed. Analysis can either be by direct sequencing (ChIP-seq) or by the hybridisation of DNA to a microarray of genomic fragments (ChIP-on-chip, or ChIP-chip). In the ensuing discussion, we focus on two studies that exemplify the issues raised by these approaches. In the first (Kim et al., 2008), a ChIP-chip approach was followed that used ES cell lines previously generated by the bio-tagging strategy (Wang et al., 2006). In the second approach (Chen et al., 2008b), ChIP-seq was used on a single ES cell line. Target proteins were collected by conventional

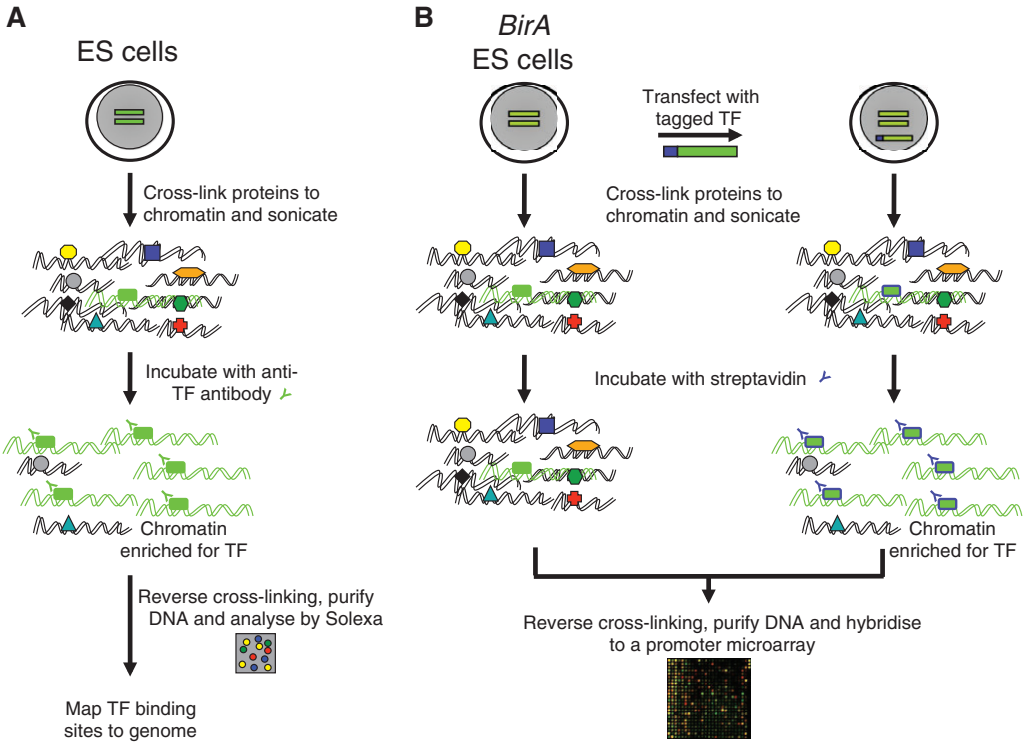

Fig. 4. Global chromatin immunoprecipitation (ChIP) methods. A comparison of the (A) ChIP-seq and (B) ChIP-chip approaches. (**A**) ES cells that express genes (green) encoding the protein to be analysed are treated with a reagent that cross-links chromatin-associated proteins to DNA. Chromatin is then prepared from the cells and sonicated to reduce the average size of the DNA fragments. Various chromatin-associated proteins (coloured shapes) are bound to the DNA (double wavy lines), including the protein of interest, which is shown in green bound to its target DNA (also in green). The sonicated chromatin is incubated with an antibody against the protein of interest and the antibody-chromatin complexes are then collected after incubation with anti-Ig immobilised on a solid support, centrifugation and washing. This enriches for both the target protein relative to other chromatin-associated proteins and its associated target DNAs (shown in green). After reversal of cross-links, the DNA is purified and analysed by Solexa sequencing and the signals corresponding to binding sites mapped to the genome (see Fig. 5). (**B**) ES cells that express the biotin ligase gene *BirA* are transfected with an expression construct that encodes a tagged version (blue) of the protein of interest (green; see Fig. 3B). Chromatin is then prepared from the *BirA* line and the derivative line expressing the protein of interest and is processed as in A, except that sonicated chromatin is incubated with streptavidin coupled to a solid phase. Following the collection and washing of streptavidin-coupled chromatin, protein-DNA cross-links are reversed and the purified DNA is hybridised to a microarray of promoter fragments that typically extend from +8 to –2 kb relative to the transcription initiation site of a subset of genes.

affinity purification using multiple antibodies, each specific for a given target protein, which could therefore be recognised in its native state (Fig. 4).

Each of these techniques offers particular advantages. Of note, the use of epitope tagging means that chromatin precipitation is not subject to the variable efficiency inherent to the use of antibodies raised against native proteins that can have significant differences in their affinity and may have differential access to epitopes present in multi-protein complexes. Conversely, transgene addition alters the dose of the transcription factor of interest. Therefore, it is important to analyse lines that minimally alter the total expression level when the aim is to isolate partners that interact with the transcription factor at endogenous expression levels. It is possible that the interaction of a protein of interest with a partner protein might be disrupted by the addition of an epitope. If chromatin localisation is dependent upon such co-binding, then a site of localisation will be missed. In practical terms, a distinct ES cell line must be generated for each protein to be analysed. The affect of this is clear from a comparison of the range of proteins studied by each approach (Table 1). In addition to Oct4, Nanog and Sox2, the proteins examined by Kim et al. (Kim et al., 2008) included the transcriptional regulators Nr0b1, Nac1 and Zfp281, which the authors had previously demonstrated to interact in solution with Nanog (Wang et al., 2006). Other proteins studied included Klf4 and c-Myc, proteins that together with Oct4 and Sox2 reprogram somatic cells to a pluripotent state (Takahashi and Yamanaka, 2006), and the Oct4 target gene *Zfp42* (zinc-finger protein 42, also known as Rex1) (Ben-Shushan et al., 1998). Chen et al. (Chen et al., 2008b) chose to examine numerous proteins that: are implicated in the control of ES cell self-renewal [Nanog, Oct4, Sox2, Esrrb and Zfx (zinc-finger protein, X-linked)]; contribute to the reprogramming of somatic cells to a pluripotent state [Klf4, c-Myc, n-Myc (Mycn)]; that regulate cell cycle progression [E2F1 (E2F transcription factor 1)]; insulate transcriptional domains [Ctcf (CCCTC-binding zinc-finger protein)]; co-activate transcription [p300 (300 kD histone acetyltransferase); Ep300]; are components of Polycomb repressive complexes [Suz12 (suppressor of zeste 12 homologue)]; or are preferentially upregulated in ES cells [Tcfcp2l1 (transcription factor CP2-like 1)]. Additional proteins examined by Chen et al. include the major mediators of the signalling pathways that lie downstream of the identified extracellular signals required to maintain ES cells in basal culture media, including BMP-induced Smad1 and LIF-

Table 1. Comparison of proteins identified in two large-scale global localisation studies

Protein	Kim et al. Bio-ChIP-chip	Chen et al. ChIP-seq
Oct4	Y	Y
Sox2	Y	Y
Nanog	Y	Y
Klf4	Y	Y
c-Myc	Y	Y
n-Myc		Y
Stat3		Y
Smad1		Y
Ctcf		Y
Nr0b1	Y	
E2F1		Y
Esrrb		Y
Nac1	Y	
p300		Y
Zfp42	Y	
Suz12		Y
Tcfcp2l1		Y
Zfp281	Y	
Zfx		Y

Data are from Kim et al. (Kim et al., 2008) and Chen et al. (Chen et al., 2008b). Y indicates that the protein was localised using that technique.
c-Myc (Myc), myelocytomatosis oncogene; Ctcf, CCCTC-binding factor; Esrrb, estrogen-related receptor beta; E2F1, E2F transcription factor 1; Klf4, Kruppel-like factor 4; n-Myc (Mycn), v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian); Nac1 (Nacc1), nucleus accumbens associated 1; Nanog, Nanog homeobox; Nr0b1, nuclear receptor subfamily 0, group B, member 1; Oct4, octamer 4; p300 (Ep300), 300 kD histone acetyltransferase; Smad1, MAD homolog 1; Sox2, SRY-box containing gene 2; Stat3, signal transducer and activator of transcription 3; Suz12, suppressor of zeste 12 homolog; Tcfcp2l1, transcription factor CP2-like 1; Zfp42 (Rex1), zinc-finger protein 42; Zfp281, zinc-finger protein 281; Zfx, zinc-finger protein X-linked.

induced Stat3. Thus, this approach allows more proteins to be examined, as multiple ES cell lines need not be derived and because commercially available antibodies can be used in the analysis.

ChIP-seq and ChIP-chip location data analysis: factors to consider

Comparisons between multiple, apparently similar ChIP experiments can produce a weaker than expected overlap, raising concerns about the reliability of the results that have been obtained (Mathur et al., 2008). The outcomes of ChIP analyses depend heavily on optimised experimental conditions. Variables include the state of the cells at the time of harvest, antibody quality, the performance of the array or sequencing method used and the data analysis methods applied. It is therefore important to consider how best to assess the biological importance of results from genome-scale ChIP experiments.

Results consistency

One way to examine the reproducibility of target predictions from ChIP experiments is to integrate results from different studies. In the large-scale studies discussed above, Kim et al. (Kim et al., 2008) used a promoter array that reports binding close to the predicted transcription start site of genes printed on the chip. Sequencing-based methods do not have this bias and report binding throughout the genome. For example, using ChIP-PET (a variant of ChIP-seq in which sequence tags are generated from both ends of cloned ChIP fragments), only 18.6% of Oct4 and 12.8% of Nanog binding sites were found within 10 kb of a predicted transcription start site (Loh et al., 2006). Therefore, the vast majority of potential binding sites might be missed in studies based on promoter arrays. Such design issues are compounded by differences in the experimental and analytical approaches used in different laboratories.

However, ChIP-seq and ChIP-chip studies can be compared if the comparison is restricted to target sequences present on the microarray used for ChIP-chip (Fig. 5). In the case of the Kim and Chen studies (Kim et al., 2008; Chen et al., 2008b), ~50% of the binding targets of c-Myc, Nanog or Sox2 identified by Chen et al. were also found by Kim et al. However, only ~20% of the Oct4 binding targets identified by Chen et al. were found by Kim et al. These overlaps are encouraging, considering the relatively simple method we have used to combine the data. It is possible that more sophisticated approaches might yield an increase in such overlaps. However, these analyses highlight the issue of using the output from an individual study, from which many putative targets might be followed erroneously.

Thresholding and prioritising ChIP targets

The purpose of ChIP analyses is to attempt to identify all the binding sites for a given transcription factor in the genome. During ChIP, the aim is to enrich DNA sequences that specifically bind a given transcription factor. In practice, a population of DNA fragments is enriched that vary in length and have random ends but which centre around the site of interest. These enriched ChIP fragments must be distinguished from the pattern of fragments obtained from controls.

In the microarray study (Kim et al., 2008), ChIP fragments were hybridised to an array of oligonucleotides designed against promoter regions. Signals from adjacent oligonucleotides generate hybridisation signals that represent putative transcription factor locations. Comparison of these putative locations with signals generated from the *BirA* control cell line is made together with standardisation of signal peak intensities with reference to the copy number, sequence composition and repeat distribution of all the oligonucleotides on the microarray (Johnson et al., 2006). Final assignment of a signal peak to a potential target gene is performed directly based on the identity of the well-characterised gene printed on the chip. In the ChIP-seq study (Chen et al., 2008b), transcription factor binding sites were identified by clustering DNA sequence outputs generated from the ChIP sample. This identifies potential transcription factor binding sites that are then compared with signal peaks computed from randomised data. Chen et al. also used an anti-GFP ChIP to account for sequences that give an elevated signal in controls (e.g. microsatellites) that cannot be accounted for by their computational model. Signal peaks from ChIP-seq experiments are commonly assigned to the nearest gene using a threshold that defines the fixed maximum distance between a binding site and an associated transcription initiation site. To avoid applying such an arbitrary threshold, Chen et al. calculated a transcription factor binding site/gene association score from the observed relationship of each transcription factor binding site with its proposed target gene and the distribution of the relevant recognition sites in randomised sequence. In this way, different association thresholds for each transcription factor between signal peak and assigned genes could be calculated.

Validation of genome-scale ChIP data

Both ChIP-chip and ChIP-seq have fantastic potential to identify transcription factor target genes and so to elucidate transcription factor networks. It is common practice to rank target lists from ChIP experiments using a variety of parameters, such as the probability of enriched binding, the number of tags that map to a particular gene, the proximity to transcription start sites or other gene regulatory

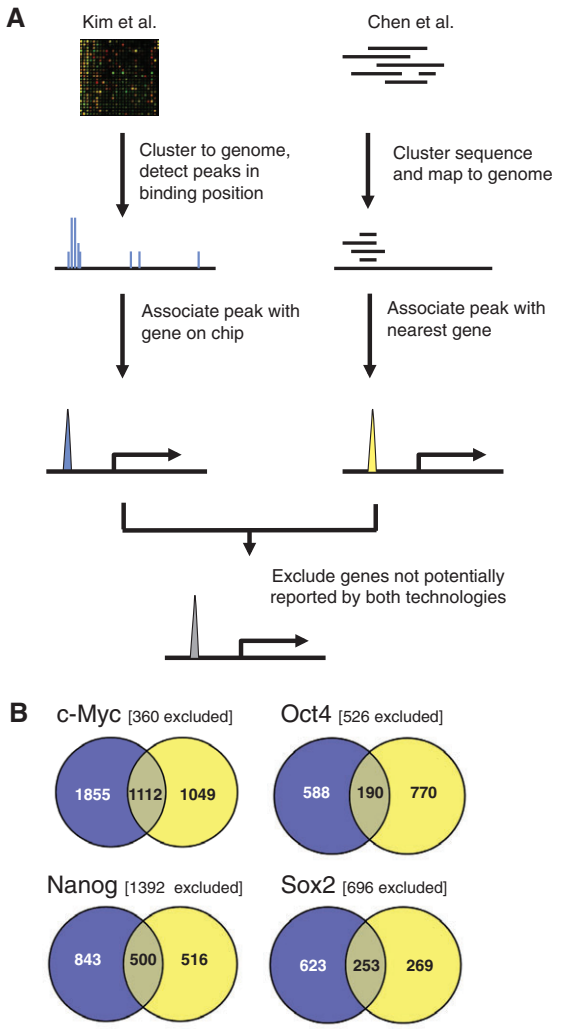

Fig. 5. Overlap in data between large-scale global localisation studies. (**A**) Workflow for ChIP-chip and ChIP-seq studies and comparison of outputs. (**B**) Overlap of target genes obtained for biotinylated c-Myc, Nanog, Oct4 and Sox2 from the studies of Kim et al. (blue) (Kim et al., 2008) and Chen et al. (yellow) (Chen et al., 2008b). To facilitate a comparison of data from the two studies, the mapping of target genes to peak binding positions was recalculated from the reported peak locations. Peaks were assigned to the nearest gene, as annotated in the Ensembl v46 mouse genomic data. Target peaks were only considered if they were in genomic regions measured by both the promoter array and the ChIP-seq study. The number of genes with ChIP-seq peaks within 10 kb of the transcription initiation site but not mapped to the microarray are shown as excluded counts.

features. In general, the relationship of these features to actual gene expression is unknown. Attempts have been made to build models to address this problem (e.g. Sharov et al., 2008) but, of course, it might be unlikely in practice that the functional activity of a given transcriptional regulator can be generally described for all genes because gene-specific factors, such as additional transcription factor binding sites, strongly influence gene expression. In addition, candidate genes for validation are almost never selected randomly but with reference to a body of existing knowledge. It is therefore impossible to use these validated genes to estimate the overall number of true positives in the candidate list.

Initial validation of genome-scale ChIP data is often performed by comparing microarray-based estimates of the transcription level of the gene to which the transcription factor binding site has been mapped. Concordance between transcription factor binding and gene expression under conditions in which the transcription factor level is modulated can then be used to identify target genes for specific transcription factors. It is instructive to reconsider the basic assumptions of this approach. Take, for example, Oct4. Depletion of Oct4 by RNAi (Loh et al., 2006) or by using the tetracycline-suppressible Oct4 cell line ZHBTc4 (Niwa et al., 2000) causes trophectodermal differentiation. Under differentiation conditions, direct targets of Oct4 may be modulated, together with indirect targets and with the associated additional epigenetic changes that occur as cells exit the ES cell state. Critically, these changes affect the accessibility of chromatin to transcription factors and might result in alterations in background binding (Rozowsky et al., 2009). This could lead to an artefactual association between differentially regulated genes and transcription factor localisation.

To avoid such spurious associations between transcription factors and non-target genes, one approach is to examine early transcriptional events, before wholesale epigenetic changes set in. In this case, the observation (Chen et al., 2008b; Kim et al., 2008; Loh et al., 2006) that many key ES cell-specific genes have multiple regulatory inputs from different transcription factors is of major importance. In a complex interconnected network, modulation of one input might not be sufficient to cause a change in transcription. An example of this can be seen in ZHBTc4 cells upon repression of *Oct4*; the levels of Sox2 and Nanog are unaffected until at least 24 hours after *Oct4* is downregulated (Sharov et al., 2008). Many other genes are modulated before these two key Oct4 targets finally begin to show reduced expression. Another approach to validation is to identify direct changes in gene expression by blocking translation (e.g. with cycloheximide) and to map these changes to ChIP binding. However, cycloheximide-treated cells are far from being in a physiological state. Moreover, a transcription factor that is bound directly to a target gene might require the translation of an unstable co-factor to become active.

These complexities mean that in practice it is difficult to connect transcriptional regulation to ChIP binding, which might partly account for the weak overlap observed between these two types of data in published studies. The proximity of a transcription factor binding site to a gene obviously does not guarantee that the binding of the transcription factor will have a functional consequence for that gene, a more distant gene or, indeed, for any gene (Sharov et al., 2008). Many other parameters need to be considered, such as local chromatin structure and the binding of co-regulators, before accurate predictions can be made. In the future, it will be desirable to develop methods for ranking ChIP binding sites based upon the likelihood that these sites will participate in gene regulation. This might emerge by calculating general 'rules' from large validated data sets and by integrating experimental data that map out the details of the local genomic context of the binding sites.

Conclusions from genome-wide transcription factor localisation studies

Each of these studies arrives at some distinct conclusions. However, the agreement between the studies is revealing. Both studies (Chen et al., 2008b; Kim et al., 2008) identify specific genomic regions that bind multiple distinct transcription factors: more than 100 promoters are occupied by seven or more transcriptional regulators and ~1000 promoters bind four or more of these proteins. In fact, the number of genes that are targets for the binding of multiple transcriptional

regulators greatly exceeds the number that would be predicted by chance. A similar increase in the number of genes bound by multiple transcription factors above that expected randomly was observed in a global localisation study of human hepatic cells (Odom et al., 2006). Therefore, the binding of multiple factors appears to have been selected for and is likely to be biologically significant. Whether the apparent co-incident binding reflects direct protein-protein interaction or mere proximity is unclear. Higher resolution analyses will be required to distinguish between these possibilities. Nevertheless, the possibility that transcription factors colocalise through direct protein-protein interactions is entirely in line with the concept of the combinatorial control of gene expression (Ptashne and Gann, 2001).

In general, it would appear that target genes bound by one or a few proteins are inactive or repressed, whereas target genes bound by four or more factors are active and become repressed upon differentiation. Kim et al. (Kim et al., 2008) used supervised clustering algorithms, which order target genes based upon their potential binding sites, to ask which target genes bound only one of the examined proteins. The largest group was genes that bound c-Myc. Moreover, most genes to which a single protein bound appeared to be repressed. It will be interesting to see how many of these genes have subunits of repressive complexes colocalised with the identified transcription factor. In this context, it is notable that subunits of the repressive NuRD complex have been reported to interact with Oct4 and Nanog (Liang et al., 2008; Wang et al., 2006). An alternative explanation for the association of lone binders with non-expressed genes is that these associations represent false positives in the localisation analyses. Finally, some lone binders might perform more interesting roles. Some transcription factors, such as the forkhead box, winged helix transcription factor FoxA1, that positively regulate transcription in differentiated cells can act as 'pioneer' factors, which are able to enter compact chromatin and make it more accessible (Cirillo et al., 2002). A Fox family member expressed in ES cells (FoxD3) can be detected by ChIP in the same region of the albumin enhancer to which FoxA1 binds in liver. This binding has been proposed to maintain unmethylated cytosines necessary to keep the site in a state compatible with normal development (Xu et al., 2007). The protein for which the evidence of a repressive function associated with lone binding looks least convincing is Oct4 (Kim et al., 2008). This could indicate that Oct4 is unique among these molecules in acting independently as a gene activator or that there are additional, unstudied Oct4 partner proteins that act in combination with Oct4 at these sites.

In addition to the redundancy of Sox proteins already mentioned (Masui et al., 2007), redundancy between other transcription factor families is likely. For example, a Klf4 binding site identified at the *Nanog* promoter (Chen et al., 2008a) has previously been suggested to be a site at which Klf5 acts (Parisi et al., 2008). Klf2, Klf4 and Klf5 have also been shown to exhibit overlapping binding site localisation (Jiang et al., 2008). However, it is also possible that such sites bind related Klf transcription factors or, indeed, Sp1 family members. In this regard, it is notable that Sp1, but not Klfs, were detected as Oct4 and Nanog partner proteins in solution (Wang et al., 2006).

A second major conclusion concerns the relationship of c-Myc to the core pluripotency factors. At the outset of these two studies (Chen et al., 2008b; Kim et al., 2008), the authors might have expected to unearth some commonality to explain the roles of c-Myc and Oct4/Sox2/Klf4 during somatic cell reprogramming. In fact the opposite is the case. Both studies found that although there are many genes that are common targets for Nanog, Oct4, Sox2 and Klf4, the genes that c-Myc localises to overlap less with these targets. c-Myc binding regions also have distinctive histone marks (Kim et al., 2008). Targets of c-Myc, along with those of Nac1 and Zfp42, are enriched for $H3K4(me)_3$, a modification that is characteristic of 'active' genes, and are depleted for the 'inactive' histone $H3K27(me)_3$ mark. By contrast, genes to which Nanog, Oct4, Sox2, Nr0b1 and Klf4 are bound are relatively enriched for both of these marks and for $H3K4(me)_3$ in particular. This profile of histone marks that are associated both with gene transcription and repression is often referred to as 'bivalency' and occurs at many promoters in ES cells (Azuara et al., 2006; Bernstein et al., 2006).

A functional distinction between the Myc binding sites and the sites to which the core pluripotency factors bind was also revealed in the Chen et al. study (Chen et al., 2008b). The transcriptional co-activator p300 predominantly localises to genomic sites that bind Nanog, Oct4 and Sox2, but not to sites that bind Myc. Moreover, the Weeder algorithm, which searches for enriched motifs in sets of sequences, detected an Oct/Sox composite sequence motif in the p300 binding site data. Genomic fragments covering 25 of the Oct4/Nanog/Sox2 clusters and eight of the Myc clusters were therefore isolated and tested for enhancer activity. All of the Oct4/Nanog/Sox2 clusters showed transcriptional activity (Fig. 6). The context in which the Oct4/Nanog/Sox2 binding sequences act is clearly important because the magnitude of the enhancer effects varied over an order of magnitude within the set of 25 test cases. Importantly, none of the Myc localisation sites had enhancer activity, further underpinning the idea that Myc functions in a way that is distinct from the other core pluripotency factors.

Several other aspects of these analyses warrant consideration. By analysing colocalisation frequency, distinct categories of colocalised binding can be seen, in addition to the major Myc and Oct4/Nanog/Sox2 clusters. Somewhat surprisingly, the protein most often colocalised with Sox2 is not Oct4, but rather Nanog (Chen et al., 2008b). It will be interesting to see whether there is a binding relationship between Nanog and Sox2 that can explain this observation. In addition, Klf4, Esrrb and Tcfcp2l1 also appear to colocalise to a significant proportion of sites to which neither Myc nor Oct4/Nanog/Sox2 bind. This might in part explain the ability of Esrrb to substitute for Klf4 during somatic cell reprogramming (Feng et al., 2009), although it should be noted that Tcfcp2l1 was tested in the same assay without success. Colocalisation is also apparent between Smad1 and Oct4/Nanog/Sox2, harking back to the protein interaction between Nanog and Smad1.

Global localisation studies during somatic cell reprogramming

Recently, the relevance of these kinds of study to reprogramming has been addressed by comparing the global localisation of Oct4, Sox2, Klf4 and c-Myc in ES cells, induced pluripotent stem (iPS) cells and partially reprogrammed cells (Sridharan et al., 2009). In partially reprogrammed cells, binding targets of Oct4, Sox2 or Klf4 that also bound c-Myc had a generally similar binding profile to those seen in ES and iPS cells. However, genes bound only by Oct4, Sox2 and Klf4 in ES and iPS cells were, in most cases, unoccupied by these proteins in partially reprogrammed cells. These results suggest that partially reprogrammed cells are locked in a state in which an additional event is necessary to facilitate the binding of pluripotency transcription factors.

This study also suggested that c-Myc acts early during reprogramming, at least in part to repress the expression of fibroblast genes, and that c-Myc might therefore function before Oct4/Sox2/Klf4. These effects of Myc expression might be a result of global alterations in chromatin, such as histone acetylation

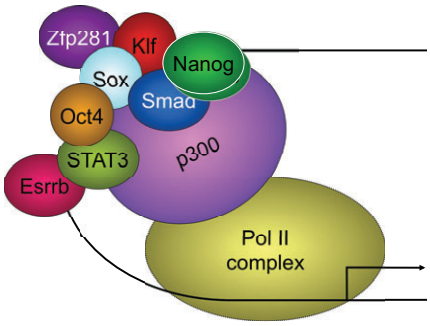

Fig. 6. Illustration of an active promoter in ES cells. Pluripotency transcription factors are shown bound to an enhancer sequence upstream of the transcription initiation site of an associated gene (indicated by the arrow). Two monomeric subunits of Nanog are illustrated to reflect the fact that Nanog is active in dimeric form (Mullin et al, 2008; Wang et al, 2008). Contact between enhancer-bound transcription factors and the RNA polymerase II (Pol II) complex machinery occurs through a bridging interaction with the p300 transcriptional co-activator.

(Knoepfler et al., 2006), and in this respect it is noteworthy that inhibition of histone deacetylases can replace c-Myc in reprogramming (Huangfu et al., 2008).

Molecules involved in reprogramming may function in an analogous fashion to pioneer factors such as FoxA1 and Gata4 (a member of a family of zinc-finger transcription factors that recognise GATA motifs), which operate on the albumin enhancer during ES cell differentiation to open up compacted nucleosomal arrays (Cirillo et al., 2002). Although c-Myc has been shown to regulate global chromatin structure in neural cells (Knoepfler et al., 2006), whether it can function in an entirely analogous fashion to FoxA1 is unknown; perhaps another pluripotency factor(s) might have this capacity.

Transcription factor co-dependency of enhancer function

Proteins binding to enhancers exhibit co-operative DNA binding (Merika and Thanos, 2001). Oct4 and Sox2 bind DNA co-operatively (Ambrosetti et al., 1997; Ambrosetti et al., 2000), but do any other pluripotency transcription factors show co-dependent DNA binding? First, consider the DNA binding of individual factors. To determine the in vivo sequence specificity of the transcription factors, sequences ±100 bp from the top 500 binding peaks for each factor were selected, repeats were masked and then searched for over-represented sequences (Chen et al., 2008b). Not unexpectedly, the sequences that underlie the DNA fragments bound by Oct4 and Sox2 were extremely similar, indicating that Oct4 and Sox2 most often act together in the relative orientation illustrated in Fig. 2. More surprising was the finding that a similar Oct4/Sox2 sequence motif underlies DNA bound by both Nanog and Smad1.

The clustering of Smad1 with Oct4, Nanog and Sox2 is interesting given that Smad1 is an intermediary in BMP signalling and that the positive effect of BMP on ES cell self-renewal can be accounted for by the Id (inhibitor of differentiation) genes. Ids are transcriptional modulators that sequester pro-neural basic helix-loop-helix proteins into non-functional complexes (Ruzinova and Benezra, 2003). In ES cells, *Id3* expression is stimulated ~6-fold by BMP treatment (Ying et al., 2003). In this regard, a Nanog/Oct4/Sox2 binding locus 1.5 kb upstream of the *Id3* gene has been described that also binds Smad1 (Chen et al., 2008b). This is significant because it suggests that some loci can bind multiple transcriptional regulators (ten in the case of *Id3*) yet remain responsive to further transcriptional activation. *Id3* might be an example of a gene that binds multiple factors without being equally responsive to all of them, but further work will be required to determine this. A related question is whether there are target genes that are bound by many of the transcriptional regulators but which remain unexpressed without recruitment of a key activating transcription factor.

In contrast to the situation for Smad1, a canonical binding site for Stat3 is found for Stat3 binding loci, suggesting that Stat3 might be more weakly associated with the Oct/Sox motif than is Smad1. Nevertheless, some Stat3 targets are co-bound by Oct4/Nanog/Sox2. Interestingly, for both Oct4/Smad1 and Oct4/Stat3 co-bound regions, the depletion of Oct4 over 2 days resulted in a reduction in Smad1 or Stat3 binding at these sites (Chen et al., 2008). If this analysis was performed prior to differentiation-induced loss of additional transcription factors, this result indicates that Stat3 and Smad1 binding to the test loci is Oct4-dependent. By contrast, perturbations in Smad1 or Stat3 did not affect Oct4 binding. The transcriptional co-activator p300 is almost exclusively localised to Oct4/Nanog/Sox2 as opposed to c-Myc targets. With the proviso mentioned above, p300 binding appears to be dependent upon Oct4/Nanog/Sox2 because depletion of any one of these proteins reduces p300 binding.

The issue of co-dependent binding has also been examined in a few gene-specific studies. Using ZHBTc4 cells, in which Oct4 protein can be eliminated by 12 hours of tetracycline treatment (Niwa et al., 2000), the binding of Esrrb to the *Nanog* promoter was found to be dependent upon Oct4 binding to a composite Oct/Sox site situated 11 bp downstream of the Esrrb binding site (van den Berg et al., 2008). In vitro studies indicate that DNA binding by Esrrb is co-operative with Oct4/Sox2. This proximity of the Esrrb binding site to the non-palindromic Oct/Sox site suggests that Esrrb makes direct, sterospecific contact with Oct4. However, the binding of Esrrb to other Oct/Sox target genes did not reveal the existence of a simple spatial relationship between DNA recognition sites (van den Berg et al., 2008). Zfp143 (zinc-finger protein 143) also binds to the proximal *Nanog* promoter, ~90 bp upstream of the Oct4 site (Chen et al., 2008a). In this case, RNAi depletion of Zfp143 caused a reduction in the localisation of Oct4 (but interestingly not of Sox2) at *Nanog* prior to the downregulation of Oct4 protein (Chen et al., 2008a).

Co-dependency of binding extends to *Xist*, the clearest example to date of a gene that is repressed by the pluripotency factors. ChIP identified an Oct4/Sox2/Nanog binding site within the first intron of *Xist*, where the pluripotency factors are responsible for repression of *Xist* and, consequently, for the activation of both X chromosomes in female ES cells (Navarro et al., 2008). In ZHBTc4 cells depleted of Oct4, *Xist* expression is massively upregulated and Nanog and Sox2 are lost from the *Xist* chromatin at a time when the *Nanog* and *Sox2* mRNAs are still expressed. By contrast, *Nanog*-null ES cells express *Xist* at ~5-fold the level of wild-type male ES cells, and Oct4 and Sox2 remain bound to the *Xist* intron. This co-dependence of binding to *Xist* is interesting in view of the fact that *Oct4* and *Sox2*, but not *Nanog*, are expressed in cells of the early female embryo in which the paternal X chromosome is inactive. Nanog may be required to facilitate Oct4/Sox2 binding to *Xist* either by recruitment or by erasure of an epigenetic imprint (Navarro et al., 2008; Navarro and Avner, 2009). There is a similarity here to the situation in partially reprogrammed iPS cells, which, like the early embryo, lack *Nanog* expression (Sridharan et al., 2009). It is possible that Nanog

facilitates full reprogramming of partially reprogrammed cells through a similar mechanism to that which operates in early embryonic cells.

These studies suggest that a hierarchy of transcription factor binding to ES cell enhancers occurs. Oct4 might nucleate the binding of the other factors, but the example of Zfp143 mentioned above indicates that more data need to be analysed before reliable rules concerning the co-dependencies of transcription factor binding to enhancers, or indeed to Myc localisation sites, can be proposed.

The global localisation studies we have discussed place Oct/Sox motifs at the centre of the pluripotency network. However, the *Sox2* deletion studies mentioned above serve as a reminder that there is still much to learn about the details of the pluripotency network. *Sox2* deletion in ES cells appears to phenocopy *Oct4* deletion (Masui et al., 2007). Consistent with this, the *Sox2* deletion phenotype can be rescued by Oct4 expression. The *Oct4* gene contains a validated Oct/Sox motif in its promoter but Oct/Sox reporters continue to be expressed in *Sox2*-deleted cells, which is likely to be due to redundancy of function between Sox proteins (Masui et al., 2007). The mechanism proposed to account for the effects of Sox2 on *Oct4* gene expression involves indirect effects mediated by Nr5a2 and Nr2f2, although this hypothesis remains to be tested. Critically, however, these considerations do not mean that the Oct/Sox motif is unimportant in the regulation of *Oct4* expression. A proper understanding will only emerge from functional studies on ES cells in which the DNA-binding motifs for each of the above-mentioned transcription factors are mutated and the consequences established.

Conclusions

Global studies of protein localisation to specific chromatin sites present us with a wealth of information on the transcriptional control of cell identity and have revealed previously unsuspected potential working relationships between transcription factors. The supposition is that many of the proteins that bind in proximity to each other are in direct contact with one another, but this remains to be established both generally and for particular loci. Additional challenges include determining the binding dependencies of individual proteins to particular loci in order to discover whether there are global rules governing all assemblies. However, it seems likely that there will be significant variation in the manner in which assemblies of transcription factors are constructed at individual loci (Fuxreiter et al., 2008), and that the details of this variation will critically determine the transcription pattern of the associated gene. Finally, the question of what determines activation versus repression remains an open issue.

Although we have focused here on studies using mouse ES cells, it will be important to understand how these findings relate to human ES cell studies. Studies on human ES cells may be best compared with studies on pluripotent mouse epiblast stem cell (EpiSC) lines, which have been established from post-implantation embryos (Brons et al., 2007; Tesar et al., 2007). ES cells and EpiSCs differ from one another in their factor requirements in vitro and in their capacity to incorporate into developing chimaeras. The recent demonstration of revertibility of EpiSCs to an ES cell state is relevant here (Guo et al., 2009). A related issue concerns the heterogeneity of the *Oct4*-expressing ES cell population with respect to its expression of pluripotency transcription factors, such as Nanog, Zfp42 and Stella (Dppa3) (Chambers et al., 2007; Hayashi et al., 2008; Toyooka et al., 2008). Understanding how transcription factor assemblies change as cells move from one pluripotent compartment to another will allow us to view how the dynamic alterations in cell phenotype that underlie developmental transitions are dictated, which will surely strengthen our ability to bend these cells to our will.

Acknowledgements

We are grateful to Sofia Morfopoulou for bioinformatic analysis and Raymond Poot for comments on the manuscript. The authors are supported by The Wellcome Trust, the Juvenile Diabetes Research Foundation, the Medical Research Council of the UK and by the EU Framework 7 project EuroSyStem.

References

Ambrosetti, D. C., Basilico, C. and Dailey, L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by sox2 and oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol. Cell. Biol.* **17**, 6321-6329.

Ambrosetti, D. C., Scholer, H. R., Dailey, L. and Basilico, C. (2000). Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the fibroblast growth factor-4 enhancer. *J. Biol. Chem.* **275**, 23387-23397.

Avilion, A. A., Nicolis, S. K., Pevny, L. H., Perez, L., Vivian, N. and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* **17**, 126-140.

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jorgensen, H. F., John, R. M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M. et al. (2006). Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8**, 532-538.

Ben-Shushan, E., Thompson, J. R., Gudas, L. J. and Bergman, Y. (1998). *Rex-1*, a gene encoding a transcription factor expressed in the early embryo, is regulated via Oct-3/4 and Oct-6 binding to an octamer site and a novel protein, Rox-1, binding to an adjacent site. *Mol. Cell. Biol.* **18**, 1866-1878.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K. et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326.

Bowles, J., Schepers, G. and Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev. Biol.* **227**, 239-255.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G. et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956.

Brons, I. G., Smithers, L. E., Trotter, M. W., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S. M., Howlett, S. K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R. A. et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* **448**, 191-195.

Chambers, I. and Smith, A. (2004). Self-renewal of teratocarcinoma and embryonic stem cells. *Oncogene* **23**, 7150-7160.

Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S. and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* **113**, 643-655.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L. and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230-1234.

Chatr-Aryamontri, A., Ceol, A., Licata, L. and Cesareni, G. (2008). Protein interactions: integration leads to belief. *Trends Biochem. Sci.* **33**, 241-242; author reply 242-243.

Chen, X., Fang, F., Liou, Y. C. and Ng, H. H. (2008a). Zfp143 regulates Nanog through modulation of Oct4 binding. *Stem Cells* **26**, 2759-2767.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J. et al. (2008b). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117.

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M. and Zaret, K. S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9**, 279-289.

Cole, M. F., Johnstone, S. E., Newman, J. J., Kagey, M. H. and Young, R. A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev.* **22**, 746-755.

de Boer, E., Rodriguez, P., Bonte, E., Krijgsveld, J., Katsantoni, E., Heck, A., Grosveld, F. and Strouboulis, J. (2003). Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. USA* **100**, 7480-7485.

Driegen, S., Ferreira, R., van Zon, A., Strouboulis, J., Jaegle, M., Grosveld, F., Philipsen, S. and Meijer, D. (2005). A generic tool for biotinylation of tagged proteins in transgenic mice. *Transgenic Res.* **14**, 477-482.

Ema, M., Mori, D., Niwa, H., Hasegawa, Y., Yamanaka, Y., Hitoshi, S., Mimura, J., Kawabe, Y., Hosoya, T., Morita, M. et al. (2008). Kruppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. *Cell Stem Cell* **3**, 555-567.

Falkner, F. G. and Zachau, H. G. (1984). Correct transcription of an immunoglobulin kappa gene requires an upstream fragment containing conserved sequence elements. *Nature* **310**, 71-74.
Feng, B., Jiang, J., Kraus, P., Ng, J. H., Heng, J. C., Chan, Y. S., Yaw, L. P., Zhang, W., Loh, Y. H., Han, J. et al. (2009). Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat. Cell Biol.* **11**, 197-203.
Fujikura, J., Yamato, E., Yonemura, S., Hosoda, K., Masui, S., Nakao, K., Miyazaki Ji, J. and Niwa, H. (2002). Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev.* **16**, 784-789.
Fuxreiter, M., Tompa, P., Simon, I., Uversky, V. N., Hansen, J. C. and Asturias, F. J. (2008). Malleable machines take shape in eukaryotic transcriptional regulation. *Nat. Chem. Biol.* **4**, 728-737.
Guo, G., Yang, J., Nichols, J., Hall, J. S., Eyres, I., Mansfield, W. and Smith, A. (2009). Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development* **136**, 1063-1069.
Hatano, S. Y., Tada, M., Kimura, H., Yamaguchi, S., Kono, T., Nakano, T., Suemori, H., Nakatsuji, N. and Tada, T. (2005). Pluripotential competence of cells associated with Nanog activity. *Mech. Dev.* **122**, 67-79.
Hayashi, K., Lopes, S. M., Tang, F. and Surani, M. A. (2008). Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. *Cell Stem Cell* **3**, 391-401.
Herr, W. and Cleary, M. A. (1995). The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. *Genes Dev.* **9**, 1679-1693.
Hochedlinger, K. and Plath, K. (2009). Epigenetic reprogramming and induced pluripotency. *Development* **136**, 509-523.
Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A. E. and Melton, D. A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. *Nat. Biotechnol.* **26**, 795-797.
Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y. and Lemischka, I. R. (2006) Dissecting self-renewal in stem cells with RNA interference. *Nature* **442**, 533-538.
Jaenisch, R. and Young, R. (2008). Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell* **132**, 567-582.
Jauch, R., Ng, C. K., Saikatendu, K. S., Stevens, R. C. and Kolatkar, P. R. (2008). Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. *J. Mol. Biol.* **376**, 758-770.
Jiang, J., Chan, Y. S., Loh, Y. H., Cai, J., Tong, G. Q., Lim, C. A., Robson, P., Zhong, S. and Ng, H. H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.* **10**, 353-360.
Johnson, W. E., Li, W., Meyer, C. A., Gottardo, R., Carroll, J. S., Brown, M. and Liu, X. S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl. Acad. Sci. USA* **103**, 12457-12462.
Kappen, C., Schughart, K. and Ruddle, F. H. (1993). Early evolutionary origin of major homeodomain sequence classes. *Genomics* **18**, 54-70.
Kim, J., Chu, J., Shen, X., Wang, J. and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell* **132**, 1049-1061.
Klemm, J. D. and Pabo, C. O. (1996). Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage. *Genes Dev.* **10**, 27-36.
Knoepfler, P. S., Zhang, X. Y., Cheng, P. F., Gafken, P. R., McMahon, S. B. and Eisenman, R. N. (2006). Myc influences global chromatin structure. *EMBO J.* **25**, 2723-2734.
Kopp, J. L., Ormsbee, B. D., Desler, M. and Rizzino, A. (2008). Small increases in the level of Sox2 trigger the differentiation of mouse embryonic stem cells. *Stem Cells* **26**, 903-911.
Kunath, T., Strumpf, D. and Rossant, J. (2004). Early trophoblast determination and stem cell maintenance in the mouse-a review. *Placenta* **25 Suppl. A**, S32-S38.
Liang, J., Wan, M., Zhang, Y., Gu, P., Xin, H., Jung, S. Y., Qin, J., Wong, J., Cooney, A. J., Liu, D. et al. (2008). Nanog and Oct4 associate with unique transcriptional repression complexes in embryonic stem cells. *Nat. Cell Biol.* **10**, 731-739.
Lints, T. J., Parsons, L. M., Hartley, L., Lyons, I. and Harvey, R. P. (1993). Nkx-2.5: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants. *Development* **119**, 419-431.
Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J. et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431-440.
Mackay, J. P., Sunde, M., Lowry, J. A., Crossley, M. and Matthews, J. M. (2007). Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32**, 530-531.
Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J. et al. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-533.
Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R., Takahashi, K., Okochi, H., Okuda, A., Matoba, R., Sharov, A. A. et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat. Cell Biol.* **9**, 625-635.
Mathur, D., Danford, T. W., Boyer, L. A., Young, R. A., Gifford, D. K. and Jaenisch, R. (2008). Analysis of the mouse embryonic stem cell regulatory networks obtained by ChIP-chip and ChIP-PET. *Genome Biol.* **9**, R126.
McDonel, P., Costello, I. and Hendrich, B. (2009). Keeping things quiet: roles of NuRD and Sin3 co-repressor complexes during mammalian development. *Int. J. Biochem. Cell Biol.* **41**, 108-116.
Merika, M. and Thanos, D. (2001). Enhanceosomes. *Curr. Opin. Genet. Dev.* **11**, 205-208.
Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M. and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* **113**, 631-642.
Mullin, N. P., Yates, A., Rowe, A., Nijmeijer, B., Colby, D., Barlow, P. N., Walkinshaw, M. D. and Chambers, I. (2008). The pluripotency rheostat Nanog functions as a dimer. *Biochem. J.* **411**, 227-231.
Navarro, P. and Avner, P. (2009). When X-inactivation meets pluripotency: an intimate rendezvous. *FEBS Lett.* **583**, 1721-1727.
Navarro, P., Chambers, I., Karwacki-Neisius, V., Chureau, C., Morey, C., Rougeulle, C. and Avner, P. (2008). Molecular coupling of Xist regulation and pluripotency. *Science* **321**, 1693-1695.
Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Scholer, H. and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* **95**, 379-391.
Niwa, H., Miyazaki, J. and Smith, A. G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat. Genet.* **24**, 372-376.
Niwa, H., Toyooka, Y., Shimosato, D., Strumpf, D., Takahashi, K., Yagi, R. and Rossant, J. (2005). Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**, 917-929.
Odom, D. T., Dowell, R. D., Jacobsen, E. S., Nekludova, L., Rolfe, P. A., Danford, T. W., Gifford, D. K., Fraenkel, E., Bell, G. I. and Young, R. A. (2006). Core transcriptional regulatory circuitry in human hepatocytes. *Mol. Syst. Biol.* **2**, 2006.0017.
Parisi, S., Passaro, F., Aloia, L., Manabe, I., Nagai, R., Pastore, L. and Russo, T. (2008). Klf5 is involved in self-renewal of mouse embryonic stem cells. *J. Cell Sci.* **121**, 2629-2634.
Parslow, T. G., Blair, D. L., Murphy, W. J. and Granner, D. K. (1984). Structure of the 5′ ends of immunoglobulin genes: a novel conserved sequence. *Proc. Natl. Acad. Sci. USA* **81**, 2650-2654.
Phillips, K. and Luisi, B. (2000). The virtuoso of versatility: POU proteins that flex to fit. *J. Mol. Biol.* **302**, 1023-1039.
Ptashne, M. and Gann, A. (2001). Transcription initiation: imposing specificity by localization. *Essays Biochem.* **37**, 1-15.
Remenyi, A., Lins, K., Nissen, L. J., Reinbold, R., Scholer, H. R. and Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* **17**, 2048-2059.
Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66-75.
Ruzinova, M. B. and Benezra, R. (2003). Id proteins in development, cell cycle and cancer. *Trends Cell Biol.* **13**, 410-418.
Sharov, A. A., Masui, S., Sharova, L. V., Piao, Y., Aiba, K., Matoba, R., Xin, L., Niwa, H. and Ko, M. S. (2008). Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics* **9**, 269.
Silva, J. and Smith, A. (2008). Capturing pluripotency. *Cell* **132**, 532-536.
Sridharan, R., Tchieu, J., Mason, M. J., Yachechko, R., Kuoy, E., Horvath, S., Zhou, Q. and Plath, K. (2009). Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136**, 364-377.
Suzuki, A., Raya, A., Kawakami, Y., Morita, M., Matsui, T., Nakashima, K., Gage, F. H., Rodriguez-Esteban, C. and Izpisua Belmonte, J. C. (2006). Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **103**, 10294-10299.
Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676.
Tesar, P. J., Chenoweth, J. G., Brook, F. A., Davies, T. J., Evans, E. P., Mack, D. L., Gardner, R. L. and McKay, R. D. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196-199.
Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K. and Niwa, H.

(2008). Identification and characterization of subpopulations in undifferentiated ES cell culture. *Development* **135**, 909-918.
van den Berg, D. L., Zhang, W., Yates, A., Engelen, E., Takacs, K., Bezstarosti, K., Demmers, J., Chambers, I. and Poot, R. A. (2008). The Estrogen Related Receptor Beta interacts with Oct4 to positively regulate Nanog gene expression. *Mol. Cell. Biol*. **28**, 5986-5995.
Wang, J., Rao, S., Chu, J., Shen, X., Levasseur, D. N., Theunissen, T. W. and Orkin, S. H. (2006). A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364-368.
Wang, J., Levasseur, D. N. and Orkin, S. H. (2008). Requirement of Nanog dimerization for stem cell self-renewal and pluripotency. *Proc. Natl. Acad. Sci. USA* **105**, 6326-6331.
Weiss, M. A. (2001). Floppy SOX: mutual induced fit in hmg (high-mobility group) box-DNA recognition. *Mol. Endocrinol.* **15**, 353-362.
Welch, G. R. (2009). The 'fuzzy' interactome. *Trends Biochem. Sci.* **34**, 1-2.
Williams, D. C., Jr, Cai, M. and Clore, G. M. (2004). Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.* **279**, 1449-1457.
Wu, Q., Chen, X., Zhang, J., Loh, Y. H., Low, T. Y., Zhang, W., Zhang, W., Sze, S. K., Lim, B. and Ng, H. H. (2006). Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells. *J. Biol. Chem.* **281**, 24090-24094.
Xu, J., Pope, S. D., Jazirehi, A. R., Attema, J. L., Papathanasiou, P., Watts, J. A., Zaret, K. S., Weissman, I. L. and Smale, S. T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **104**, 12377-12382.
Yamanaka, S. (2007). Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell Stem Cell* **1**, 39-49.
Ying, Q. L., Nichols, J., Chambers, I. and Smith, A. (2003). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115**, 281-292.
Zhao, S., Nichols, J., Smith, A. G. and Li, M. (2004). SoxB transcription factors specify neuroectodermal lineage choice in ES cells. *Mol. Cell. Neurosci.* **27**, 332-342.